Planetary Data System

Archive Preparation Guide (APG)

DRAFT

September 13, 2004

Version 0.040913

Planetary Data System
National Aeronautics and Space Administration

JPL

Jet Propulsion Laboratory
Pasadena, California

JPL D-xxxxx

Planetary Data System

Archive Preparation Guide (APG)

September 13, 2004

Version 0.040913

Approved:

Laverne Hall — Date
PDS Project Manager

Reta Beebe — Date
PDS Project Scientist

Bill Knopf — Date
PDS Program Executive

Jet Propulsion Laboratory
Pasadena, California

JPL D-xxxxx

CHANGE LOG

Revision	Date	Description	Author
Start Draft	2004-06-01	Initial outline	Gordon/Simpson
Draft Release 1	2004-06-15	First Draft	R. Joyner
Draft New Outline	2004-07-09	Outline approved by MC	Working group
0.040712	2004-07-12	Second Draft	R. Joyner
0.040721ras	2004-07-31	Wholesale slash and burn	R. Simpson
0.040820	2004-08-20	Revised per telecon 2004-08-09	R. Simpson
0.040820a	2004-08-30	Small changes to 3.2.2; rewrite remainder of Section 3	R. Simpson
0.040906	2004-09-06	Rewrite Sections 4, 5	R. Simpson
0.040909	2004-09-09	Editorial improvements	R. Simpson
0.040913	2004-09-13	Updated Editorial improvements	R. Simpson

TABLE OF CONTENTS

1.0 INTRODUCTION

You are part of an investigation, instrument, or mission team whose planetary exploration proposal has recently been accepted. Now you need a real plan for real archive data and a real schedule. What do you do?

1.1 Purpose

This *Archive Preparation Guide* (APG) provides a step-by-step, “cookbook” approach to preparing data for submission to NASA's Planetary Data System (PDS).

1.2 Audience

The target for the APG is the data manager responsible for creating and delivering archival data from a planetary science instrument or investigation to PDS. The audience will also include software professionals designing the archive and operators involved in production. To a lesser extent it may include the principal investigator or team leader.

The document provides fairly detailed insight into the PDS data archiving process and may be of interest to scientists proposing for NASA funding, NASA managers, archivists, and other data management professionals. Proposers should consult the PDS *Proposer's Archiving Guide* (PAG) [3] before reading this document.

We encourage you to identify and work closely with a PDS Discipline Node (DN) "mentor" during the archiving process; mentors will need to be familiar with the contents of this document.

1.3 Scope

The APG is written primarily for use in a flight project (mission) context; but many of the principles will apply to other situations, such as archiving results from a data analysis program or recovering data from pre-PDS experiments.

The archiving process can vary significantly from one mission to another — and for a variety of reasons. Of these, mission size often has the biggest impact. The general approach here is to discuss a typical large mission and then identify adjustments which make sense in other cases.

The APG is written to be consistent with version 3.x of the PDS *Standards Reference* [1]. Examples are limited to a small subset of the possibilities allowed; but these are intended to be the most common and popular. In fact, we explicitly recommend that you *not* exercise all of the options allowed. Consult with your PDS mentor if you stray much beyond the recommendations here, as PDS is considering narrowing the range of choices to simplify future operations.

1.4 Document Overview

Producing an archive generally proceeds in three steps. After some brief comments on archiving philosophy, we describe each of these steps in some detail:

- Planning and Design
- Development and Testing
- Data Production, Distribution, and Maintenance

1.5 Applicable Documents

The following documents may be obtained from the PDS Operator (phone number??) (pds_operator@jpl.nasa.gov), or downloaded from the indicated URL:

[1] Planetary Data System Standards Reference, August 1, 2003, Version 3.6, JPL D-7669, Part 2. (http://pds.jpl.nasa.gov/stdref)

[2] Planetary Science Data Dictionary Document, August 28, 2002, JPL D-7116, Rev. E. (http://pds.jpl.nasa.gov/psdd.pdf)

[3] Planetary Data System (PDS) Proposer's Archive Guide (PAG), Version 1.0, June 15, 2003, JPL D-26359. (http://pds.jpl.nasa.gov/documents/documents.html)

[4] Planetary Data System (PDS) Archive Life Cycle (ALC), **TBD**

1.6 Additional Resources

The following may also be of interest.

[101] Planetary Data System (PDS) web site – starting point for accessing all PDS information

http://pds.jpl.nasa.gov

~~[102] Planetary Data System Internal web site – as this site is under password protection, you will need to contact the Discipline Node for access.~~

~~**http://pds.jpl.nasa.gov/internal**~~

[103] *Archive Preparation Guide* web site – this document and related materials.

http://pds.jpl.nasa.gov/documents/apg/

[104] PDS Personnel / Contact List – table of PDS personnel and other contacts. These individuals are your most useful resource – contact them early and often.

http://pds.jpl.nasa.gov/data_services/contacts.html

[105] PDS Interactive Phonebook – interactive search tool for PDS personnel.

http://pds.jpl.nasa.gov/tools/phonebook.cfm

[106] Quick-Start Introduction to PDS Archiving – concise introduction to archiving with the PDS (how is this different from or the same as APG??)

http://pds.jpl.nasa.gov/documents/qs/index.html

[107] PDS Data Dictionary Lookup tool – interactive search tool for searching the keywords and objects that comprise a PDS label.

http://pds.jpl.nasa.gov/tools/data_dictionary_lookup.cfm

[108] FAQ – list of frequently asked questions covering production of an archive.

TBD

2.0 OVERVIEW OF THE ARCHIVING PROCESS

PDS archives and distributes documented data to the planetary science community. Its functions include validation, ingestion, and distribution of data at 'discipline nodes' (DN's) specializing in certain scientific disciplines and/or technical skills. DN's, their host institutions, and their areas of expertise are listed in Appendix F.

Data handled by PDS must be understandable, be in formats that future scientists will find easy to use, and follow standards for organization and content that facilitate machine-assisted correlative science across missions and science disciplines. PDS requires that submitted data meet published standards regarding format, content, and documentation [1, 2]. A range of formats has been defined, but 'tables' and 'images' account for the majority of products. 'Data products' are organized logically into 'data sets,' which are sometimes organized into 'data set collections'. Data may be written to 'volumes,' 'volume sets,' and 'volume series' on physical media such as CD-ROMs and DVDs.

At an early stage, mission and instrument personnel should define the data products intended for archive, group them logically into data sets, estimate their volume and generation rate, and negotiate a preliminary delivery procedure with PDS. Adequate documentation for both understanding and using the data is critical to each archive; mission planners should be mindful of the archive documentation requirement as they develop explanatory materials for other purposes. All archive submissions to PDS are peer-reviewed by scientists and data engineers to ensure that PDS standards have been met, that the archive is complete, and that the data are useful. Review usually occurs at several stages for mission archiving.

Once data are fully integrated into PDS, they may be retrieved through electronic queries over the internet or, by special request, on physical media such as CD-ROM. PDS provides copies of all accepted data sets to the National Space Science Data Center (NSSDC), which serves as both the PDS 'deep archive' and the official distribution point for requests originating outside the NASA planetary research community.

2.1 Elements of a Good Archive

A useful mission archive, as required by most NRA’s and AO's, includes raw data from each instrument, data calibrated in physical units, and derived products based on further processing of the data. To the extent these can be generated routinely based on expectations before data collection begins, they are called 'standard products.' The results of more advanced processing (especially processing developed in response to early findings) and/or combinations of different data such as maps, overlays, and comparative tables are called 'special products.' Special products may also be included in an archive—and may be among the most useful to other scientists—but they are not usually required.

The archive contains sufficient documentation of the mission, the instrument(s), and the calibrations that a scientist of a future generation could intelligently use the data and, if appropriate, even recalibrate the data. The archive includes complete information about the geometry relevant to the observations (e.g., spacecraft position and orientation relative to the target). It includes indices to help in searching the data and catalog files which can be ingested into the PDS database.

Finally, the archive uses a straightforward organization and a small number of widely recognized, non-proprietary formats. Data should be described as they exist in the archive files; specialized software (whether included in the archive or not) should not be required to display or manipulate them. And, although PDS recognizes and allows a wide range of formats, the 'good' archive should use formats that are both simple and popular wherever possible.

2.2 PDS Concepts

PDS jargon is sometimes an obstacle to the new archivist—particularly since some terms have very specific meanings within PDS. This section introduces key terms and concepts as they are used within the PDS. They are presented here in order to facilitate a clear and concise understanding of the material which follows. Some of the same terms also appear in Appendix B, which provides more technical definitions.

2.2.1 Data - Logical Building Blocks

A PDS archive includes primary data and supporting material organized into a hierarchical structure. For purposes of this document, we use the terms 'primary data' or 'science data' to mean the primary output from instruments making scientific observations or measurements. Within the rubric of 'primary data' we also include the results from the processing of instrument output—calibrated data, resampled or gridded data, maps, etc. Tables of summary output from many observations by many instruments may also be 'primary data.' The supporting material—also called 'ancillary data'—is needed to use and understand the primary data and may include calibrations, geometry information, documents, software, indexes, and files used for cataloging.

The hierarchy is shown in Figure 2.2.1; the individual terms are described in more detail below.

- **data** - PDS data (both primary and ancillary) exist in the form of digital files.

- **data object** - Data may be organized into about 30 different PDS-recognized formats called 'objects,' such as TABLE and IMAGE. A file can contain one or more objects, some of which may be nested. For example, a file may contain two TABLE objects, one containing instrument settings and the second containing measurements. By definition, each of these TABLE objects must include one or more COLUMN objects.

- **label** - Data objects are described by collections of keyword=value statements called labels. If the label is in the same file as its object(s), it is called **attached**; if the label and objects are in separate files, the label is called **detached**. See below for more discussion of keywords and their values.

- **data product** - The smallest locatable unit of data within PDS is the data product—a label and the one or more objects that it describes.

- **data set** - A data set is a collection of related primary data products and the ancillary data products needed for their understanding and use. The primary data products typically (but not always) are a series of observations from an instrument which have been processed in the same manner. Each data set has a unique DATA_SET_ID and DATA_SET_NAME; each data product has a unique PRODUCT_ID within the data set.

- **data set collection** - Related data sets may be grouped into a data set collection - for example, all of the data sets (raw, calibrated, and processed) from a single instrument.

2.2.2 Data - Units of Storage

Data need a home; we identify three levels of storage, which are also shown in Figure 2.2.2.

- **volume** - Data sets are stored on physical or logical media within an archive; one physical or logical unit of such storage is called a volume—for example, a single compact disk, magnetic tape, or disk partition. Each volume is identified by a unique VOLUME_ID and VOLUME_NAME. Depending on capacity of the volume relative to size of the data set(s), a single volume may hold several data sets or only part of one data set.

- **volume set** - a collection of related volumes is called a volume set and is identified by a unique VOLUME_SET_ID and VOLUME_SET_NAME. The volumes in a set often come from a single source. For example, the 150 CD-ROM volumes of raw data from the magnetometer on the Geronimo mission might be identified in their labels by

  ```
  VOLUME_SET_NAME = "GERONIMO MAGNETOMETER RAW DATA"
  ```

- **volume series** - a collection of related volume sets is called a volume series and is identified by a unique VOLUME_SERIES_NAME. The sets in a volume series usually come from different sources. For example, all of the Geronimo Mission data (raw and reduced from all instruments) could be a volume series; or, all of the data from a single planet (from all missions) could form a series.

  ```
  VOLUME_SERIES_NAME = "MISSIONS TO NEPTUNE"
  ```

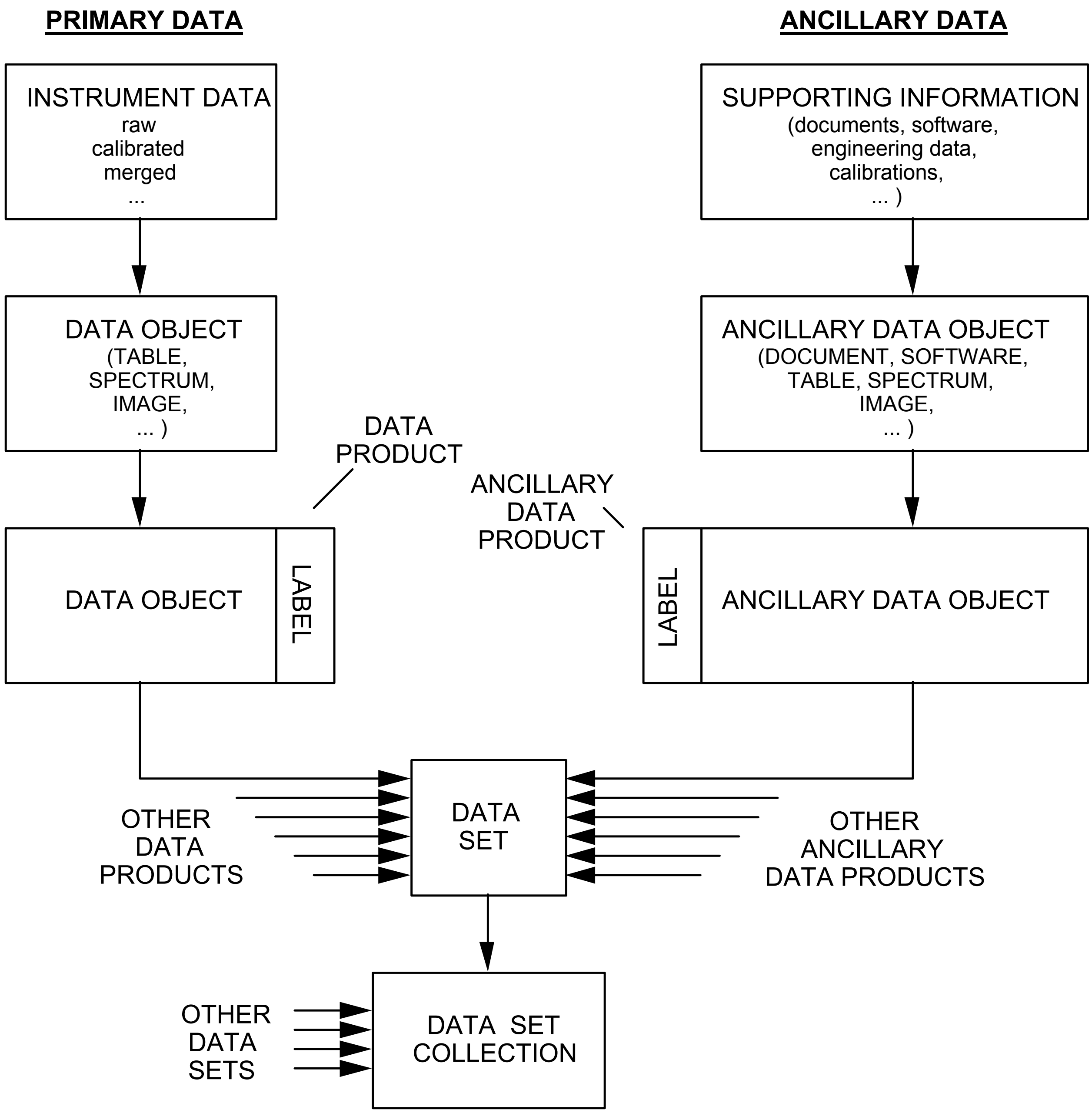

Figure 2.2.1. Hierarchy showing relationships among files, objects, data products, labels, data sets, and data set collections. Primary data originate with an instrument; ancillary data are everything else needed to use or understand the primary data.

2.2.3 Keywords and Values

The label is important because it describes both the content and the format of its data product. The information in a label is used to catalog each data product so that it can be located directly and uniquely by its storage address or found as part of a search (where are all magnetometer raw data products from all missions collected on 13 May 2004?). Labels are composed of keyword=value pairs in text format.

- **keyword** - Keywords identify characteristics of data files and objects. For example, they may specify the version of the PDS Standards under which the data product was defined, list physical characteristics of the file (e.g., number of records and bytes per record), and identify and describe the objects (object types, file names, data set, etc.). Each keyword must be listed in the *Planetary Science Data Dictionary* (PDSDD) [2] or a local data dictionary (LDD). If used, LDD's are generally set up and maintained by mission personnel.

- **value (or keyword-value)** - A keyword-value can be a single value, an ordered sequence of values, or an unordered set of values. In the examples below, keywords are on the left, values on the right. Each pair is sometimes called an **attribute**.

```
PDS_VERSION_ID          = PDS3
RECORD_TYPE             = FIXED_LENGTH
RECORD_BYTES            = 80
FILE_NAME               = "TESTDATA.TAB"
```

- **standard value** – Some keywords may take on only a limited number of pre-defined values, known as "standard values." In the examples above, `RECORD_TYPE` is such a keyword; its values are limited to:

```
FIXED_LENGTH
STREAM
VARIABLE_LENGTH
UNDEFINED
```

The current set of PDS keywords can be found in the PDSDD [2] or on-line using the PDS Data Dictionary Lookup tool [107]. If no existing keyword meets your needs, there are several options; contact your PDS/DN.

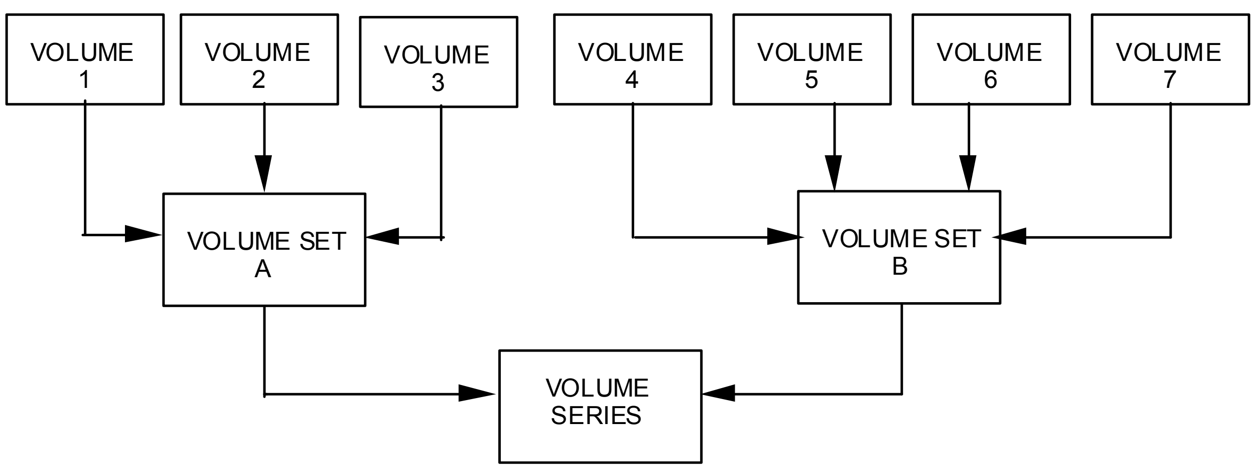

Figure 2.2.2. Relationships among volumes, volume sets, and volume series. Each volume is a unit of physical or logical storage within an archive; volume sets are aggregations of volumes, and volume series are aggregations of volume sets.

2.2.4 Documents and other Supporting Material

In order for planetary science data to be useful to those not directly involved in their collection, supporting information—*e.g.*, documents—must be included in the archive. Such documents are expected to meet not only the PDS label and format requirements, but also the structural, grammatical and lexical requirements of a refereed journal submission. Documents submitted for archiving which contain spelling errors, poor grammar or illogical organization will be rejected and may ultimately lead to rejection of the submitted data.

2.2.5 Format Options

Anticipated use of the science data will often dictate the archiving format—*e.g.*, TABLE, IMAGE, SPECTRUM, etc. For example, most spacecraft image data sets will be composed of single image data products, each defined as an IMAGE object. Most data that can be visualized in rows and columns are well-suited to becoming TABLE objects, the SPECTRUM being a special case of the TABLE. The SPREADSHEET object may provide more efficient storage when fields are highly irregular and/or many values are missing. TABLE, SPECTRUM, and SPREADSHEET objects can be read into many database applications. Although the PDS has defined a range of data formats, 'tables' and 'images' account for the majority of products. If in doubt about the best object to choose for your data, contact your PDS/DN.

3.0 ARCHIVE PLANNING / DESIGN

This chapter provides an overview of archive planning and design, which includes:

- Making an inventory of your data holdings (or imagining your future holdings)
- Understanding and applying PDS archiving concepts
- Designing the archive
- Prototyping the archive

We will illustrate these steps using common, practical examples. PDS allows a wide range of implementations; but the Standards can sometimes be strict. If you deviate significantly from the examples here, we suggest you stay in close contact with your PDS/DN mentor(s) to minimize false steps.

3.1 Inventory Data Holdings / Production Plans

Taking an 'inventory' will give you an idea of both how big your archiving task is and also how many useful parts of the process are already in place. To get started, ask yourself the following questions:

- What data are (or will be) available?
    - Data types
    - Data rate
    - Total data volume
- What documentation is (will be) available?
- What calibration is (will be) available?
- What are the steps in processing the data?
- How and when will the data be peer reviewed and validated?
- Where does archiving fit within the data production pipeline?
- Is anything obviously missing in the inventory?

All of these questions have been asked before—by both beginners and experienced data producers. Many answers fall directly out of your mission or instrument data management plan; but there may be slight differences when you view them as part of an archiving activity.

Get a sense of the data flow. What goes into your data processing pipeline? Where should you tap into the pipeline to pick off the archival files? Do NOT invent an entirely new pipe for archiving; that wastes time and effort and jeopardizes your quality control. You want the products that go into the archive to be the same products that have been thoroughly examined and tested by your team's scientists.

3.2 Preliminary Design

A common mistake of first time archivists is to design the archive solely by reading PDS documentation. Use the experience of your PDS/DN mentor at this stage. Someone has probably produced an archive like yours previously; there is no harm in copying a good model. In fact, there may be advantages to a future scientist if your archive looks like others in the PDS holdings; it will be easier to compare data from different times, missions, and instruments. PDS personnel can point you toward examples of previous work.

Once you have a good sense of your own data flow, you need to define the products in PDS terms (objects, products, data sets, etc., as introduced in Section 2). The principal activities are:

- Defining the data products
- Defining the data sets
- Developing file and directory naming algorithms
- Identifying pipeline production issues
- Defining the data validation process

Some of these activities will be done in parallel; assigning file and directory names as you define data products is often convenient. In many cases there may be several iterations at each step. A single naming algorithm that works on all of your files can pay handsome dividends; but you may have to experiment with many different file types before you get all the bugs out.

We expand on these five activities in the sections below.

3.2.1 Define the Data Products

Data products should be defined based on science requirements—in many cases the definitions will have already been made by your team before you start thinking about archiving. With raw data, the products are usually obvious—the records or files spewed out by the instrument. For partially or fully processed data, there are more choices; you can select parameters that will be more useful (*e.g.*, calibrated radiance) or drop others that seem unimportant (status flags of heaters that only operate during eclipses). If the products have already been defined, you may still have some options regarding file sizes and formats. Note, however, that the more you can use an existing conceptual pipeline to deliver archival products, the simpler and more successful your archiving task will be. Mainstream products will have been well-scrubbed by the time they go to archive *and* you will have to implement fewer unique steps to get them there.

If data product definition has been left largely to you, the archivist, consider the following.

3.2.1.1 Data Product Usage

Typically, data products are defined based on factors such as their utility in supporting science investigations of the team, the probable way in which the data will be accessed, and the expected frequency of access. What does the team need in order to meet its science goals? What processing schemes have been employed in the past—either by your team or others? What types of information are presented at scientific and project meetings? Will any of the results be fed back into future planning? If so, what and how? Is your team centrally located? If not, are there concerns about delivering data to remote sites either because of the volume or poor communications links? Does your data appear in typically small or large blocks? If small, are there logical ways to aggregate several small blocks into a larger product? If large, does it make sense to subdivide the blocks into smaller pieces without affecting the science use?

3.2.1.2 Selecting Objects

Once you have determined the 'logical' definition of your data products, you need to format them according to PDS Standards. About 30 PDS objects have been defined including TABLE, IMAGE, SPECTRUM, and COLUMN; see Appendix A in [1] for details.

TABLE is a natural choice for numerical data which can be easily visualized in rows and columns. The PDS TABLE object accommodates both binary and ASCII tables; the latter is encouraged for all but the most massive files or if the data originate from another source and there is danger that conversion might lead to file corruption. Note that COLUMN is also a PDS object; each PDS TABLE must include at least one COLUMN object. Such nesting allows you to build complex products from more primitive modules. The example in Figures 3.2.1.2a,b shows how a simple list of hourly temperatures can be defined as a TABLE with four COLUMN objects. The data object (which we assume is a single, separate file) is given in Figure 3.2.1.2a; the object definition, which will be part of its detached label, follows in Figure 3.2.1.2b. Note that the value of ROW_BYTES includes delimiters but that the values of START_BYTE and BYTES do not and that the double quotes in Column 3 are considered to be the delimiters of a character string.

```
12:00:00,  43200,"F9",  19.4
13:00:00,  46800,"F9",  20.1
14:00:00,  50400,"F9",  22.9
15:00:01,  54001,"B4",  23.0
16:00:00,  57600,"F9",  22.5
```

Figure 3.2.1.2a. Example TABLE object.

```
OBJECT = TABLE
  INTERCHANGE_FORMAT = ASCII
  ROWS               = 5
  COLUMNS            = 4
  ROW_BYTES          = 30
  DESCRIPTION        = "Hourly temperatures outside lab.
    Column delimited by commas. Rows delimited by ASCII
    carriage-return line-feed pairs."
  OBJECT             = COLUMN
    NAME               = "UTC"
    COLUMN_NUMBER      = 1
    DATA_TYPE          = TIME
    START_BYTE         = 1
    BYTES              = 8
    DESCRIPTION        = "Time in UTC on 2004-05-12."
  END_OBJECT         = COLUMN
  OBJECT             = COLUMN
    NAME               = "TIME IN SECONDS"
    COLUMN_NUMBER      = 2
    DATA_TYPE          = ASCII_INTEGER
    START_BYTE         = 10
    BYTES              = 7
    FORMAT             = "I7"
    UNIT               = SECOND
    DESCRIPTION        = "UTC converted to seconds."
  END_OBJECT         = COLUMN
  OBJECT             = COLUMN
    NAME               = "SENSOR ID"
    COLUMN_NUMBER      = 3
    DATA_TYPE          = CHARACTER
    START_BYTE         = 19
    BYTES              = 2
    FORMAT             = "A2"
    DESCRIPTION        = "Thermometer identifier. First
      character gives location: B, F, S = back, front,
      side door. Second indicates model number:
      4 = Temp-o-dat model 451; 9 = Ice9 Fridge Sensor"
  END_OBJECT         = COLUMN
  OBJECT             = COLUMN
    NAME               = TEMPERATURE
    COLUMN_NUMBER      = 4
    DATA_TYPE          = ASCII_REAL
    START_BYTE         = 23
    BYTES              = 6
    FORMAT             = "F6.1"
    UNIT               = CELSIUS
    DESCRIPTION        = "Temperature outside the lab."
  END_OBJECT         = COLUMN
END_OBJECT         = TABLE
```

Figure 3.2.1.2b. Example TABLE object definition using nested COLUMN objects.

Structurally speaking, SPECTRUM is simply a TABLE object defined for spectral data. SPREADSHEET bears some similarities to TABLE except that it uses FIELD objects instead of COLUMN's. Each FIELD has only a *maximum* byte count, allowing some fields to be empty. This makes SPREADSHEET attractive in terms of storage if many values are missing.

The IMAGE object is flexible in the sense that both ASCII and binary samples (pixels) can be accommodated. The image is specified by keywords which give the number of lines and samples and the properties of the latter. It is possible to assign part of the beginning and/or end of each line for storage of non-image data. Figure 3.2.1.2c shows an example of the simplest IMAGE object definition.

```
OBJECT          = IMAGE
  LINES           = 960
  LINE_SAMPLES    = 956
  SAMPLE_TYPE     = UNSIGNED_INTEGER
  SAMPLE_BITS     = 8
END_OBJECT        = IMAGE
```

Figure 3.2.1.2c. Example IMAGE object definition

3.2.1.3 Estimate File Sizes and Data Flow

Once data products have been defined, estimates of the file sizes need to be made. This is so you can predict the rate of data flowing into the archive and begin thinking about delivery schedules. Writing a CD every few months is not a heavy work load; but missions are being planned now that bring into question the viability of CD and DVD as transfer media. The workload in writing, validating, and distributing 40000 unique DVD volumes per year becomes almost unmanageable given today's technology.

Assuming you are not on the "viability frontier" as described above, estimate the number of archival products of each type that will be generated by your team within some time interval, multiply by the size of the respective products, and sum the results. Divide the sum by the capacity of your preferred storage media (e.g. 650 Mbytes per CD) to obtain the production rate in media-units per time interval. Think about whether this is reasonable and, if not, whether there are alternatives.

It is common within missions to stage data deliveries on quarterly (for example) time scales. That is, you hold your archival CD's and deliver them to PDS only at the end of March, June, September, and December. Not all missions operate this way, nor do all teams within single missions.

3.2.2 Define Data Sets and Data Set Collections:

The data set is a logical grouping of data products. Defining the scope of a data set is often done before, or concurrently with, defining the individual data products. All of the raw data from your instrument and the necessary supporting information could be one data set; the calibrated data could form the basis for a second data set, and the processed data a third. Observation type, discipline, target, or time may also be used to discriminate among data sets.

If your mission includes several targets in succession—for example, two asteroid flybys followed by orbiting a comet (a hypothetical AFCR mission)—you may want to define separate data sets for each body. If only a limited amount of data is collected at each asteroid, you may choose to combine the raw, calibrated, and processed data into a single data set for each asteroid. On the other hand, if the orbit around your comet lasts for several years, you may find it easier to manage those data by separating the files into data sets corresponding to mission phases (or years). You could have both raw and calibrated data sets for each of the AFCR primary, extended, and extended-2 phases at the comet. Since future scientists unfamiliar with the details of your mission may not care about its mission phases, you could group all of the reduced data into a single data set.

Data sets may be combined into data set collections. All of the data sets from one instrument in the AFCR mission could be considered a data set collection, for example.

Definition of both data sets and data set collections is, in many respects, arbitrary; you need a reason for the grouping, but many reasons (or combinations of reasons) will suffice. In a mission context, data sets and data set collections are often defined in conjunction with other data producers, and in consultation with PDS advisors. Together you jointly select mission and instrument acronyms and build keyword-values for DATA_SET_ID and DATA_SET_NAME. For example

```
DATA_SET_ID     = "AFCR-A-ISS-1-VESTA-V1.0"
DATA_SET_NAME   = "AFCR RAW IMAGE DATA FROM VESTA V1.0"
```

See the "Data Set/Data Set Collection Contents and Naming" chapter in [1] for details.

One example of a data set collection is the Pre-Magellan CD-ROM containing selected Earth-based radar data from Venus, the Moon, Mercury, and Mars; Pioneer Venus radar data; airborne radar images of Earth; and line of sight acceleration data derived from tracking the Pioneer Venus Orbiter and Viking Orbiter 2. The unifying theme in assembling this collection was its possible use in helping to understand data collected from Magellan.

3.2.3 Develop Naming Conventions and Algorithms for Files and Directories

File and directory names should be intuitive—they should allow the user to navigate from one part of the data set to another without getting lost. It may be convenient to organize data (especially raw data) chronologically—for example, directories named by year and month and files named by day, hour, minute, and second.

The naming convention adopted should result in names that uniquely identify the directories and files within an easily recognized pattern that can be used by both humans and machines. For example, if T810111.DAT contains data from January 11, 1981, then use of the wildcard T8101* selects all of the data from January.

Numeric dates in year-month-day order are preferred since FEB_2004 will appear before JAN_2003 and 02_01_2004.DAT will appear before 2004_02_01.DAT in most computer listings. Names based on observation sequence numbers will also generally result in a time ordering of directories and/or files. Image or map data can be named efficiently by including target coordinates; spectral data may be organized using names which include wavelength.

PDS recommends that directories contain no more than 256 files or subdirectories — approximately the number that would fit conveniently onto a single computer screen. If more exist, then adding a new level of subdirectories is preferred—up to a limit of eight total directories in depth. By using subdirectory names efficiently in conjunction with file names, both can be kept short.

> NB: Each data product must have a unique PRODUCT_ID within its data set. If directory and file names are kept short, then some of the directory path may need to be preserved in the PRODUCT_ID to prevent duplication. For example, if images at two latitudes, but covering the same longitudes, are stored with the same file names, the uniqueness can be preserved as follows:

```
FILE_NAME  = "E340_350.IMG"          /*IN DIRECTORY N10_20*/
PRODUCT_ID = "N10_20_E340_350.IMG"

FILE_NAME  = "E340_350.IMG"          /*IN DIRECTORY N40_50*/
PRODUCT_ID = "N40_50_E340_350.IMG"
```

PDS allows directory and file names to have as many as 31 characters. In the examples immediately above, the longer form could have been adopted for both FILE_NAME and PRODUCT_ID. This helps to ensure that all file names, which are more readily visible to both humans and computers, are unique and may reduce confusion.

See the "File Specification and Naming" and "Directory Types and Naming" chapters in [1] for further information.

3.2.4 Identify Pipeline Production Issues

Whether they use the term or not, instrument teams generally develop a 'pipeline' for handling mission data; to do otherwise risks consistency in processing and possible loss of data. The pipeline begins with data collection (as from a telemetry stream) and ends with generation of 'standard' products, which may then be distributed to team members for analysis, fed back into planning for future observations, or both. The standard products may be data at any level of processing starting from telemetry after artifacts have been removed to fully calibrated and registered products. The point is that automation is required to handle the data flow—with appropriate human monitoring for quality assurance.

As the archivist, you should take advantage of this pipeline for several reasons. Except for a few ancillary products (documents, for example) it will provide most of the input you will need for your archive. More importantly, it will provide these products in the same (or very nearly the same) form as your team will see them; thus 'bugs' in the archive will be detected and reported by team members in the course of their regular science analysis.

The pipeline, as developed to meet team needs, may not meet all of your archiving requirements; modifying the pipeline *before production begins* can meet these additional needs while retaining (and possibly enhancing) its original advantages. For example, your pipeline may not include generation of labels. Adding labels forces you to define objects and adopt a logical file naming scheme; having a single name for each file rather than different names for operations and archiving can pay significant dividends when you try to track down problems in the archive. The labels themselves provide your team members automatically with much of the identifying information they will need later when preparing for publication or presentations.

Visualizing the pipeline gives you a second, more practical look at your data and its flow than your original inventory. Review the questions originally posed in Section 3.1, or use the ones listed below which cover some of the same ground.

- How and where do you define data products?
- How and where do you create labels?
- What are the appropriate levels of processing for data going to the archive?
- How do you aggregate the data into data sets and archival volumes?
- How much data will be archived?
- What types of calibration files and algorithms are needed?
- What delivery time(s) are appropriate?

3.2.5 Identify Data Validation Issues

Even if you fully exploit the pipeline, there will be a few validation steps that will be unique to the archive. Only in unusual circumstances will your team members see the actual archive volumes, for example; in most circumstances, the data products will be distributed immediately for team use while creation of the volumes will lag behind. Experience tells us that team members do not willingly go back through archive volumes simply to provide validation.

PDS has tools which can assist you in these steps where the team has gone missing. For example, the PDS Volume Verifier (*VV*) scans each archive volume to check that you have all of the expected pieces in place. The Label Verifier Tool (*lvtool*), Table Browser Tool (*tbtool*), and *NASAVIEW* provide convenient ways for you to check that individual products meet PDS Standards while you are debugging the software that creates them. *lvtool* is one of the modules run automatically by the Volume Verifier, so every label receives this validation when you run *VV*. For the others, it is usually considered sufficient that you validate the pipeline software (including after each software upgrade) and make occasional spot checks during production.

In a mission, you will often be asked to demonstrate in a review that you have developed a reasonable archive design; you will also be asked to circulate example products and volumes for review prior to production. Depending on the convening authority, your review may be conducted by others in the mission, outside science reviewers, and/or PDS staff. Criteria will include adequacy of the archive in terms of content and documentation and adherence to PDS Standards. Once production begins, there will likely be a thorough validation of the first few volumes by outside reviewers, then occasional spot checks to ensure that errors have not crept into the process.

3.3 Prototyping

Although prototyping may have seemed implicit in our discussions to this point, we list the task explicitly here to emphasize its importance. A smoothly running archive delivery system requires that its many pieces work together efficiently; each time a human intervenes there is both a loss of time and the possible introduction of errors. This is not to say, of course, that every pipeline leading to archival products will be error-free; but your chances of introducing random errors (spelling mistakes, missing elements, etc.) goes down sharply as you increase the amount of automation. Careful prototyping shows what each step in the pipeline must do and helps you identify the issues at the interfaces between steps.

We identify the key steps in prototyping here, then expand on each in the following subsections. In most cases you will want software to generate files automatically, so we emphasize that aspect of prototyping; but you can also assemble products by hand—simply to see what they look like.

- Primary Data
  - Instrument data objects
  - Labels

- Ancillary Data (and their labels)
  - Calibration data (including geometry data)
  - Catalog files
  - Index files
  - Documentation, software, etc

3.3.1 Primary Data

At this point, you should have identified the data object that will describe your data product; if you have several different primary products, you need an object for each. The data files themselves should have the formats expected of PDS TABLE's, SPREADSHEET's, IMAGE's, etc.; if not, you need to modify your pipeline software so that the primary data products meet those format specifications.

Using examples from previous archives (and with advice from your PDS mentor), you can then produce labels for each product type. In addition to the object definitions already roughed out (*e.g.*, Figures 3.2.1.2b and 3.2.1.2c) you will need keywords (and values) to:

- identify the version of PDS Standards you are using (PDS_VERSION_ID)
- identify the data product (*e.g.*, PRODUCT_ID and DATA_SET_ID)
- characterize the file physically (*e.g.*, RECORD_TYPE and RECORD_BYTES)

- describe data object contents (*e.g.*, FILTER_NAME and/or OFFSET_MODE_ID)
- point to the data object (*e.g.*, pointer ^IMAGE)

The chapter "Data Product Labels" in [1] has more information on label construction.

If prototyping software, write it to find keyword-values within the data files or from other readily accessible sources; label-making software that requires long lists of command line arguments will be clumsy and more prone to operator errors. Observation time tags often accompany even the most primitive raw data files; convert these to meaningful dates and times for the label. On the other hand, if you have no easy way to find the surface target coordinates of a raw image frame, perhaps that label information should be deferred to a later step in processing where you have reconstructed the viewing geometry. When done, use *lvtool* on your prototypes; then run them past your PDS/DN mentor, your team members, and anyone else who will look.

In general, PDS encourages you to define a DATA directory within each archive volume. Subdirectories under DATA then hold the products themselves. For large data sets, you may wish to order the products chronologically, placing the earliest products on the first volumes and later products on subsequent volumes. Your PDS/DN mentor can provide guidance on this organization based on the number of files you expect, their sizes, and the rate at which they are generated.

3.3.2 Calibration Files and Algorithms

All calibration information needed for interpretation of primary data (including data needed to reconstruct the viewing geometry) must be produced, documented, and archived coincident with the science data that they describe. Notes explaining the calibration procedure need to be drafted and included in the archive. Some teams will have software which actually allows a user to calibrate the data. Calibration software may be included in the archive; but such software is considered 'supplemental' and is not acceptable as a substitute for the required calibration algorithms.

It is particularly important that you address calibration data while prototyping the main archive (or earlier!) since some instruments have significant volumes of pre-launch calibration data that must not be lost in the excitement of mission operations. The suite of calibration data, algorithms, and (possibly) software will allow recovery of the full scientific value of the returned data and facilitate the correlation of data taken at different times, by different instruments, and by other spacecraft, observing platforms, and missions.

It is common to store calibration data within a CALIB directory on each archive volume, repeating the files on each volume if they represent only a small fraction of the volume capacity. Large volumes of pre-launch calibration data may be handled more conveniently by using one or more separate volumes. Your PDS/DN mentor can provide guidance on the appropriate way to archive calibration data.

If your mission (and your instrument) uses the SPICE system developed and supported by the PDS Navigation Ancillary Information Facility (NAIF), you may opt to join in a common archive of SPICE files produced by NAIF. The NAIF archive has the advantage of including all SPICE software and documentation as well as the data files. Note, however, that the NAIF archive may be dynamic—as spacecraft trajectory and attitude reconstructions are improved, the set of files in the NAIF collection may be updated. Your archive needs to document the steps followed by *your* team which, at some point, may include archaic versions of these files. If your calibrated and/or reduced product labels point toward the SPICE files actually used and these can be found in the common archive, the requirements for archiving geometry information may be met. If the files your team used are not in the common archive, you may need to include a team-specific SPICE collection in your own archive. You may also choose to archive SPICE files simply for the convenience of having them locally. Check with your PDS/DN mentor, who should be up-to-date on plans for the common SPICE archive, for guidance if you are uncertain about the geometry archiving requirements in your case. Note that there is a small set of SPICE-specific keywords that may be useful in constructing labels for SPICE products.

SPICE and other geometry data are usually stored in the GEOMETRY directory on archive volumes; this directory may be subdivided into other directories for spacecraft and planetary ephemerides, attitude files, clock conversions, etc. As with the CALIB directory, it is convenient if the geometry files needed to analyze primary data are included within the same volume; this may require duplication of files from volume to volume.

Don't forget that calibration and geometry data must be formatted as PDS objects—just as primary data are—and that each object must be part of a data product described by a label. Some calibration and geometry data naturally fit within the definition of TABLE, IMAGE, and other objects we have mentioned in connection with primary data. But you will also find a need for the DOCUMENT, TEXT, and (possibly) SOFTWARE objects.

3.3.3 Catalog Files

All data submitted to PDS must be accompanied by a set of catalog files which briefly describe the mission, instrument host (that is, the spacecraft or other facility within which the instrument operates), instrument, and data set. A fifth catalog file identifies key personnel associated with the instrument, data set, and archiving task. A sixth file lists references cited in the first four files or which might otherwise be useful to a scientist working with the data at some future time. These six files usually appear in the CATALOG directory of each archive volume. Some archives include additional catalog files, such as for targets of the instrument; these are usually considered optional but may be encouraged or required by some missions.

Catalog files are written in a structured format but so as to be easily read by humans. Your responsibility (at the instrument level) is limited to INST.CAT, DATASET.CAT, and PERSON.CAT. MISSION.CAT and INSTHOST.CAT should be written and maintained by the flight project; you only need to download the current versions and add them to each of your archive volumes. A baseline REF.CAT is drafted by the project—including the references cited in MISSION.CAT and INSTHOST.CAT; you then need to add the references you cite in INST.CAT and DATASET.CAT. Some projects, in conjunction with PDS, maintain a full REF.CAT—to which you contribute; this has the advantage that all references have been checked against and entered into the PDS data base before you deliver any volumes. Otherwise, you run the risk of submitting REF.CAT entries that conflict with entries already in the data base. This most often occurs when you draft a REF.CAT entry with a REFERENCE_KEY_ID that has already been used. A less common problem, and one that is more difficult to detect, is when two or more archivists submit REF.CAT entries for the same paper but with different REFERENCE_KEY_ID values.

PDS maintains a set of catalog file templates. These provide the outline structure expected; you just fill in the text. Or, you may find it convenient to take an existing catalog file and use that as a model, swapping in your own details where they differ.

Most catalog files evolve over a mission. For example, the flight project will update the MISSION.CAT file as new mission phases are defined. MISSION.CAT also documents important mission events, which may or may not have been predicted. A series of 'safing' incidents could affect data collection; the dates when these occurred will be of interest to future users of the data sets. Your own DATASET.CAT file has a section on "Data Coverage and Quality" which gives a very coarse summary of where and when you collected data and whether anomalous conditions existed.

Rather than submit complete catalog files with your initial data delivery, you may choose to submit a set of partially completed skeleton files. Once the data become available to users through PDS, complete files should be entered in the PDS data base; but the evolutionary nature of these files argues for some flexibility in their completeness during the time when data are still being acquired. Contact your PDS/DN mentor and mission archive coordinator for guidelines.

A seventh file (VOLDESC.CAT) appears in the root of each archive volume. Although technically a 'catalog' file, we have deliberately omitted it from the discussion here since its construction and use is somewhat different.

3.3.4 Index Files

Index files tell users where to find data products. There is an INDEX.TAB file in each volume; when there is more than one volume, there is a CUMINDEX.TAB for all volumes in the volume set. The INDEX_TABLE is a distinct PDS object; but, as with other TABLE objects, each file must have a label in which the INDEX_TABLE object is defined.

For each product, the index entry gives the VOLUME_ID, the path to the product, the PRODUCT_ID, the DATA_SET_ID, and the PRODUCT_CREATION_TIME. PDS recommends that other 'fingerprints' be included. For instance, associating latitude and longitude ranges or start and stop times with each product will help future users in searches for related files.

PDS requires that all primary data objects be listed in INDEX.TAB (and CUMINDEX.TAB) and recommends that a wide variety of ancillary data products also be listed. But the extent to which data producers include ancillary data varies from archive to archive. A good rule-of-thumb is to include all data products that might be used in numerical calculations; this leaves things like DOCUMENT and SOFTWARE objects out of INDEX.TAB. For guidance on current practice, consult your PDS/DN mentor.

3.3.5 Documentation

Documentation provided with the data set should be sufficient that a future scientist could understand how the data were collected and what they mean. To the extent that such discussion might help with the understanding, an explanation of why the data were collected and what might have been done differently could also be valuable. The role of the ancillary data should also be clear. If calibration data are included, a description of the calibration procedure is required. If some of the data have been 'processed,' the steps in processing need to be described.

In the mission context, it is common for science teams to write papers explaining their goals and the measurements that will be carried out to achieve those. Copyright protection may preclude such papers from the archive; but some publishers have made exceptions. Check with your PDS/DN mentor regarding possibilities for gaining permission to include such documents.

The format of the data included in the archive must be described clearly. Full PDS labels should provide this information; but it is common for many missions to require one or more Software Interface Specifications (SIS's) or Interface Control Documents (ICD's) for both individual data products and also for volumes. Appendix D provides a possible outline for the latter.

Documents usually appear in the DOCUMENT directory of a volume. Simple text files can appear as TEXT objects with attached labels; these are among the simplest products in an archive. More complicated documents can be built around an ASCII text core file, supporting files with graphics, and a label which wraps all of the components into a single DOCUMENT product. Including special formats such as Microsoft Word and Adobe PDF is also encouraged since these may be more readable in the short term. The human-readable ASCII core file is *required* in all cases, however, since ASCII text is expected to be viable over a longer time span than any of the specialized formats.

3.4 Archive Planning and Design Checklist

The following checklist is provided as a summary of the steps outlined in the Archive Planning / Design section. The checklist can be used as a means of ensuring the steps towards creating a successful archive are complete.

	SECTION	
☐	1.0	Review PAG
☐	1.0	Review DMAP and ICD / SIS
☐	1.0	Review PDS Standards & PSDD
☐	2.1	Become familiar with the elements of a good archive
☐	2.2	Become familiar with PDS concepts and terms
☐	3.1	Inventory your data holdings - review the mission or instrument data management plan - identify areas ripe for discussion
☐	3.2	Establish a preliminary design for the archive - identify similar PDS holdings suitable as a model
☐	3.2.1	Define Data Products (parameters)
☐	3.2.1.1	Establish data product usage (factors)
☐	3.2.1.2	Select appropriate PDS object(s) for your data products
☐	3.2.1.3	Estimate data product file sizes - predict the rate at which the data will become available - predict the rate at which you will archive data - predict delivery schedules
☐	3.2.2	Define the Data Set(s) and Data Set Collection(s) - define the scope of the data sets - define how data sets are grouped into collections - identify conventions (for both names and identifiers)
☐	3.2.3	Adopt a logical scheme for organizing your directories and data files - define naming conventions (for both directories and data product files) - ensure each data product has a unique PRODUCT_ID within the data set
☐	3.2.4	Identify pipeline production issues - identify any tools / processes that will be specific to your pipeline
☐	3.2.5	Identify data validation issues - pay attention to validation steps that are unique to your archive - become familiar with the PDS validation tools
☐	3.3	Plan out a prototype of your data production pipeline - keep your PDS/DN in the loop / ask for suggestions

	SECTION	
☐	3.3.1	Draft sample labels for each data product - keep your PDS/DN in the loop / ask for suggestions
☐	3.3.2	Draft sample calibration files / algorithms that will be used for interpreting the primary data - keep your PDS/DN in the loop / ask for suggestions
☐	3.3.3	Draft catalog files (as complete as possible) - keep your PDS/DN in the loop / ask for suggestions
☐	3.3.4	Draft sample index files - keep your PDS/DN in the loop / ask for suggestions
☐	3.3.5	Draft sample documentation files - keep your PDS/DN in the loop / ask for suggestions

4.0 DEVELOPMENT AND TESTING

By 'development and testing' we mean the adaptation of your team's data processing pipeline to generate PDS-compliant archival data products and volumes. After you have designed and prototyped your products, setting up a system to produce and assemble them automatically into volumes is the next step. Then you need to deliver them (after appropriate review) to PDS—a step which may be beyond your team's original conceptual pipeline.

4.1 Development

The main elements of a typical data processing pipeline are shown in Figure 4.1. The data are collected, inspected, cleaned up, and stored in the first stage; calibrated in the second; and processed to standard products in the third. If the flow is designed to support archiving, raw primary and ancillary data can be formatted as PDS objects, named, and labeled before entering the team data base. As derivative files are created later in the pipeline, they are also named and labeled to PDS standards. The archiving step then becomes little more than bookkeeping—staging the new files until enough have been collected to assemble a new volume, creating and validating the volume, and delivering it to PDS.

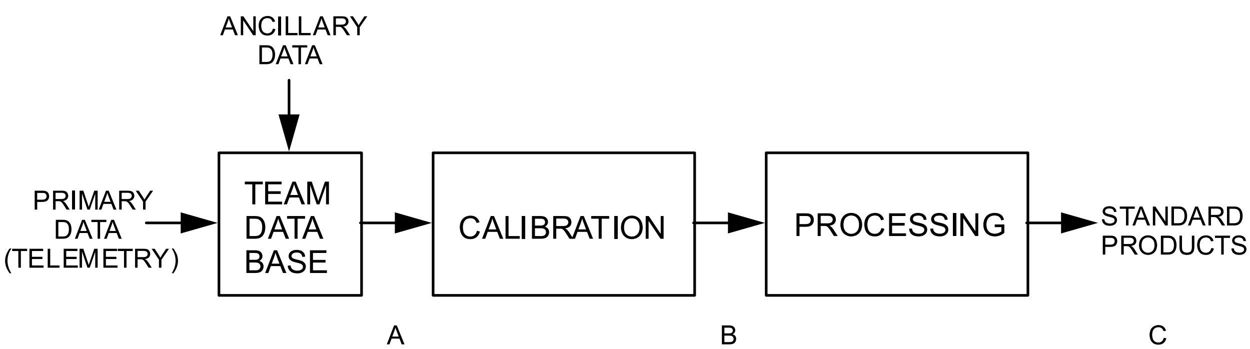

Figure 4.1. Simplified pipeline for mission processing. Data from the instrument and other sources are stored (with labels) in the team data base, then used as input to a calibration stage. Later generic 'processing' yields standard data products. Files for the archive can be obtained at points A, B, and C; in NASA terminology, these correspond generally to data of Levels 0, 1, and 2 (see Appendix E).

Since PDS and future users of the data need to understand the steps leading to the products at steps A, B, and C (Figure 4.1), we encourage you to:

- keep your PDS/DN mentor in the loop on what the pipeline will do
- keep your PDS/DN mentor in the loop on what the pipeline will not do
- ensure that documentation is updated in parallel with changes to the pipeline and its software.

4.1.1 Label Generation

Your PDS/DN mentor can offer suggestions on automating the process for generating labels, including the use of PDS tools and commercial software. Certain types of label generation may benefit from using the PDS Label Library. Other PDS tools available specifically for label generation include:

- TAB2LAB: a table-to-label generator that produces labels from a table of ASCII data—e.g., an index table.
- FILE2LAB: a file-to-label generator that may be used to produce labels from existing ASCII data files.

In some cases, the label becomes comparable to or exceeds the length of the product itself. You can simplify the label by abstracting its repeated portions into 'format' (FMT extension) files which are stored only once; the individual labels then reference the FMT file(s) through pointers. Your PDS/DN mentor can help with this and other techniques to simplify your work.

lvtool is an important adjunct to automating your label generation; use it frequently to make sure you have not introduced anomalies into your labels.

4.1.2 Volume Production

The "Volume Organization and Naming" chapter of [1] tells you what goes into a volume. Let's assume for the remainder of this section that you are part of a large mission and have one raw data set spread over many volumes. We'll go through the steps to produce a volume of raw data. There may be second and third data sets for calibrated and reduced data, respectively; they can be handled in the same way.

Your data base (Figure 4.1) can help in volume production by telling you how much data has accumulated since you made your last volume. Or, you may simply keep the new files in a separate area where you can watch them grow until the new volume has been created and the new files have been merged with the old in the data base. If you have 600 MB in new files (for CD volumes), it's time to master the next volume.

In addition to the new primary data files from your instrument and new ancillary files of geometry and other data, there will be a set of 'static' files that go into every volume—for example, your CALIBRATION, DOCUMENT, and SOFTWARE directories. These are files that do not change—or change so infrequently that you can imagine hand-editing them when updates are needed. You should keep these in a convenient, but safe, location from which they can be copied.

PDS also requires the CATALOG directory and its files. You may need to update the "Data Coverage and Quality" section in DATASET.CAT; otherwise, CATALOG should be essentially static across many volumes.

AAREADME.TXT, ERRATA.TXT, and VOLDESC.CAT are files in the root of each volume; they generally need small updates for each volume. You will probably be making continual updates to ERRATA.TXT as you discover minor errors and 'features' in the data set; these updates occur asynchronously with volume production. The other two files will be customized for each new volume.

INDEX.TAB and its label need to be created for each new volume; the new INDEX.TAB is then appended to the previous CUMINDEX.TAB, and CUMINDEX.LBL is updated accordingly. Archivists have developed various pieces of software to assist you in generating index files; contact your PDS/DN mentor for suggestions and assistance. Note that INDEX_TABLE is a separately defined object within PDS, having its own requirements.

4.1.3 Volume Validation

There are two parts to volume validation. First is confirmation that the volume meets archiving standards. This can be done automatically (albeit not perfectly) using tools available from PDS. These check that the volume has all required components, that all files are properly recorded in INDEX.TAB (and CUMINDEX.TAB), etc. They also call other PDS utilities (such as *lvtool*) to make sure that labels are correct, that all references citied in other CATALOG files actually appear in REF.CAT, and that line lengths and delimiters in text files meet PDS recommendations. Ask your PDS/DN mentor for help; then get the Volume Verifier installed and running on your system.

The second part of volume validation is confirmation of the scientific integrity of the contents. This is more difficult since it is partly subjective. If your pipeline is designed so that team members are using the same products for their research as you include in the archive, your job is vastly simpler. If the path to the archive differs significantly—for example, you generate the products long after the science team has studied other versions—you and your team will have to develop a separate validation process. Scientific integrity of the volume also includes checking that documentation is sufficient to explain and use the data; if your DOCUMENT, SOFTWARE, and CALIBRATION directories are complete for the first volume, all you need do is double check that the same files have been incorporated into subsequent volumes. If your instrument or processing change, you will, of course, need to keep the ancillary files current. Be careful, however, not to lose information important at the beginning of the mission by removing it from later versions of the files—unless you carry forward all versions of the documentation.

If your team has written one or more SIS/ICD's and/or a Data Management Plan, they may include sections on validation. Check that each step mentioned in each SIS/ICD has been implemented in your archiving pipeline. Make sure that each person with identified responsibilities in validation is ready to perform and understands those steps.

Implementing data validation is often an iterative process. Both the procedures themselves and the software written to perform certain steps will likely require a number of updates as your validation becomes both more comprehensive and also more automated.

4.2 Testing

Testing is done at two levels—the team level, where you assemble example products (based on simulated data, if possible) into sample volumes; and at the mission/PDS level, where those volumes are reviewed for compliance with PDS Standards and adequacy of documentation and other ancillary information. In some circumstances, it may be desirable to conduct at least the part of the PDS review concerned with archive design *before* you have made significant investments in software development; check with your PDS/DN mentor for guidance.

4.2.1 Team Testing

4.2.1.1 Data Product Simulation

Simulation is important because it gives you concrete examples of data products to test data flow and interfaces; your team may already be using simulated data to test its pipeline. Sample volumes of products generated from simulated data may be used to test archive interfaces with your mission and/or PDS. Real data (as from pre-launch bench tests) can also be helpful, but they may not be as robust in exercising processing steps that involve expected observing geometries. Tests with simulated data products may also provide insight on desirable enhancements to products (*e.g.*, additional keywords) that you may have missed in your initial design and prototyping.

4.2.1.2 Pipeline

As early as possible, you should get the data production pipeline into place. Once the pipeline is working, you can input simulated raw data (as from your instrument) and simulated files from other sources and test all of the steps, including generation of sample archive volumes. Check that the pipeline operates on the simulated data at each step as you expect.

4.2.1.3 Sample Volumes of Archival Data

The sample volumes should be "complete" in the sense that they are representative of what a final "archive" volume will contain. They should contain examples of each data product using simulated or test data as available. It is important that the structure of the volume be accurately represented so that all automated validation steps can be tested. Once this step has been accomplished, you can focus on the second part of validation—checking the scientific integrity of the volume. This means getting real science data into the primary files and fleshing out the documentation and other ancillary information. If your mission has a "cruise" phase, you may be able to use those data for more realistic testing.

4.2.2 Testing at the Mission and/or PDS Level

Every archival data product (and volume) must meet standards for format and content set by PDS. Whether that has been accomplished is determined by a peer-review committee. There are usually two levels of peer review in a mission context—an evaluation of the archive design, including sample products and volumes; and a review of mission data as volumes are delivered to PDS. Although the objectives are similar in both, the design review is often more thorough. Assuming that liens identified in the design review have been successfully resolved, the delivery review panel only spot checks that volumes continue to meet the design specifications. In the remainder of this section, we focus on the archive design review.

There are four general areas in which the archive is reviewed:

- Compliance with PDS Standards
- Compliance with the applicable Data Management Archive Plan (DMAP), Software Interface Specifications (SIS'S), and/or Interface Control Documents (ICD's)
- Scientific merit of data (are these the proper data to be archived?)
- Usability of the data (appropriate formats; completeness of data set, including ancillary data; comprehensive documentation)

Steps in the review process include:

- Establishing a review committee
- Conducting the review
- Completing the review (including documenting and resolving liens)

4.2.2.1 Establishing a Peer Review Committee

When you are ready for your archive design review, contact your PDS/DN mentor who will recruit the review committee. In large missions, a mission archive manager may handle the arrangements and dictate a schedule; but it is PDS that is most concerned about the outcome. Members of the committee typically include the manager of your mentoring Discipline Node, the data preparer (possibly you) or Principal Investigator (PI), one or more data suppliers (*e.g.*, someone from the Spacecraft Team or Ground Data System), one or more outside scientists from the planetary community capable of critically reviewing the data, and a PDS data engineer. The PDS/DN manager typically chairs the review committee.

4.2.2.2 Conducting the Review

You will be responsible for providing copies of the design documents (DMAP, SIS's, and or ICD's) and sample data products or volume(s) (*e.g.*, CD-WO or DVD media) to each member of the review committee. Electronic access to them over the Internet is usually considered acceptable.

The review committee chair will provide instructions on how the review will be conducted. Sometimes you will be asked to walk the committee through the design, products, and volumes during a plenary session, which may be conducted as a face-to-face meeting, teleconference, or both. In other cases, the chair may ask committee members to provide written comments (*e.g.*, e-mail) on the design and sample products in the four areas mentioned above before the panel convenes. In unusual circumstances, the chair may ask the committee members to submit their reviews in advance, and the plenary will be your opportunity to respond to their comments.

If your archive includes special software for displaying or manipulating the data, it is expected that committee members will evaluate the software both functionally and as a long-term component of the archive. Reviewers should also evaluate documentation (including labels, calibration files, and CATALOG files): do they explain what the data mean and how to use them, and will they be intelligible to a scientist 10, 20, or 50 years in the future?

4.2.2.3 Completing the Review

At the end of the peer-review meeting the chair will summarize the results of the review in a formal memo. The review committee is responsible for making a recommendation on whether:

- The design documents and sample data products/volumes passed peer-review and development can proceed without additional review.
- The material did not pass peer-review and must undergo an incremental (delta) review in certain specified areas.
- The material did not pass peer-review and must undergo a full peer-review.
- The material did not pass peer-review and does not merit a second peer-review.

In addition to an overall recommendation, the chair will compile a list of "liens" — questions about or requests for change in the archive design or sample products/volumes. The liens must be resolved before the archive can be accepted by PDS. Note, however, that you are not obligated to accept all requests for change. You may offer a rebuttal to one or more liens; it is then up to the chair and PDS as to how the lien should be "resolved." In extreme cases, where the peer-review committee collectively agrees that a product does not meet archiving requirements and the team is unable or unwilling to resolve the liens, then PDS may elect to “safe store” the data (*i.e.*, the data is not deemed archive quality, but may be preserved and distributed). "Safing" data has been allowed only rarely; it is most appropriate when a data 'restoration' improves the quality of old data but does not have the resources to respond to all liens resulting from a review.

If the design and the sample products/volumes have been approved by the review committee, you should complete your development and testing at the team level and move to the “Data Production” stage.

4.3 Archive Development & Testing Checklist

The following checklist is provided as a summary of the steps outlined in the Archive Development and Testing section. The checklist can be used as a means of ensuring the steps towards creating a successful archive are complete.

	SECTION	
☐	4.0	Begin pulling the pieces together to set up a system to automatically assemble data products into volumes
☐	4.1	Review the elements of a typical data production pipeline
☐	4.1.1	Plan out how to automate label production - identify PDS tools to help automate the process - identify any tools that will be specific to your label production process - contact your PDS/DN for suggestions
☐	4.1.2	Plan out how to automate volume production - identify any tools that will be specific to your volume production process
☐	4.1.3	Plan out how to validate each volume in production - validation process ensures compliance with archiving standards - validation process ensures confirmation of scientific integrity - review SIS/ICD validation requirements
☐	4.2	Plan out the two levels of testing - keep your PDS/DN in the loop / ask for suggestions
☐	4.2.1.1	Produce sample volume(s) using simulated data - validate data flow and production interfaces
☐	4.2.1.2	Ensure that each step of the data production pipeline operates as you expect
☐	4.2.1.3	Ensure the sample volumes are representative of what a “final” archive volume will contain
☐	4.2.2	Ensure sample volumes meet standards for format and content - review the PDS policy governing quality assurance
☐	4.2.2.1	PDS/DN will establish peer-review committee
☐	4.2.2.2	Provide copies of the SIS/ICD and sample “archive” volumes to the peer-reviewers - participate in the peer-review - be prepared to demonstrate usability of the data

	SECTION	
☐	4.2.2.3	The peer-review chairperson will document any review liens - you are responsible for post peer-review lien resolution - may require follow-on peer-review

5.0 DATA PRODUCTION / DISTRIBUTION / MAINTENANCE

This chapter provides an overview of archive production, distribution, and maintenance. Maintenance includes reprocessing, which may be required for several reasons.

5.1 Data Production

Data Production entails running the data production pipeline, validating the archive volumes, and transferring them to PDS. The means of transfer and the schedule should be defined in the DMAP, usually with concurrence of the mission.

Data Production also includes peer review of products and volumes delivered to PDS. Rather than the comprehensive review of design and sample products/volumes described in the previous chapter, the review of delivered products is usually conducted using automated tools. It checks for format and PDS compliance on many, if not all, volumes. Often the first few volumes are examined very carefully to ensure that they follow the (previously approved) SIS or ICD; but then only spot checks are made to confirm science integrity on succeeding volumes. Only if significant problems are discovered or if the delivered products and volumes deviate from the designs approved in the earlier review is the flow of data to PDS interrupted.

5.2 Data Distribution

Once the data have been received, PDS is responsible for releasing them to the science community and the public. In PDS “data release” means adding data products to an online repository, ensuring that the data are accessible, and then announcing their availability. PDS operates a "subscription manager" service which allows potentially interested scientists to sign up for automatic notifications of data releases.

Data products may be distributed by PDS through a DN, the Central Node, or a Data Node working in conjunction with a PDS/DN. The organization responsible for distributing each data product is identified in the DMAP. Data are officially distributed to the general public (anyone who is not a NASA-funded planetary scientist) through the National Space Science Data Center; the DMAP specifies how NSSDC obtains its copies of the archive.

Depending on the data set and urgency of the request, users may obtain products from PDS either by downloading files from the repository or on physical media such as CD-ROM.

5.3 Maintenance

Maintenance is the set of processes by which you respond to errors or updates in data products/volumes. These include:

- Reprocessing—if the team determines that there has been an error in previous processing or that refined algorithms will produce better results
- Withdrawal—if the team determines that there has been an error in processing and that previously released products cannot be repaired
- Update ERRATA.TXT—if an error of relatively little consequence has been identified and the cost of repair significantly outweighs any benefit, the team may simply note the problem in ERRATA.TXT
- Update file—Some files can be updated with little impact on the archive as a whole. For example, corrected typographical errors in documentation or updates of "Data Coverage and Quality" in DATASET.CAT often do not warrant even the little attention provided by a note in ERRATA.TXT.

Any changes in the production and/or validation of the data products/volumes should be reflected in the design documents.

~~Note: Excessive re-processing of data products may require the PDS to negotiate new areas of responsibility and funding for the additional time and effort to re-validate and re-release archive products. (RAS: I think I know what is meant here; is there a better way to say it?)~~

NB: Extensive reprocessing may violate assumptions made by PDS when it agreed to your mission's (or instrument's) archive plan. If you expect significant fractions of your archive to be delivered more than once, you should make that clear in the archive plan (by amendment, if necessary).

5.4 Data Production / Distribution / Maintenance Checklist

The following checklist is provided as a summary of the steps outlined in the Data Production / Distribution / Maintenance section. The checklist can be used as a means of ensuring the steps towards creating a successful archive are complete.

	SECTION	
☐	5.0	Become familiar with the archive production, distribution, and maintenance steps
☐	5.1	Begin running the data production pipeline, validating the archive volumes, and transferring the volumes to the PDS
☐	5.2	Ensure the PDS has “released” the products, the data are accessible, and an announcement of the availability of the products was dispatched
☐	5.3	Be prepared to respond to errors or updates in either data production or volume production

APPENDIX A ACRONYMS

The table below lists acronyms and abbreviations used in this document.

AO	Announcement of Opportunity
APDTP	Archive Policy Data Transfer Plan
AP	Archive Plan
ASCII	American Standard Code for Information Interchange
CDR	Critical Design Review
CD-ROM	Compact Disc—Read-only Memory
CN	Central Node
CN/DE	Central Node / Data Engineer
CODMAC	Committee On Data Management And Computation
COTS	Commercial off-the-shelf
CSR	Concept Study Report
DAWG	Data Archive Working Group
DE	Data Engineer
DIM	Digital Image Model
DMAP	Data Management and Archiving Plan
DMP	Data Management Plan
DN	Discipline Node
DTM	Digital Terrain Model
DVD	Digital Versatile Disc
ESA	European Space Agency
FTE	Full Time Employee or Full Time Equivalent
GDS	Ground Data System
HQ	Headquarters
ICD	Interface Control Document
ID	Interdisciplinary Scientist
IDSMA	Interdisciplinary Scientist for Data Management and Archive
ISO	International Standards Organization
MRR	Mission Readiness Review
NAIF	Navigation and Ancillary Information
NASA	National Aeronautics and Space Administration
MAC	Mission Archive Coordinator
MOU	Memorandum of Understanding
NAIF	Navigation Ancillary Information System
NRA	NASA Research Announcement
NSSDC	National Space Science Data Center
ODL	Object Description Language
Ops	Operations
PAG	Proposer's Archive Guide
PDR	Preliminary Design Review
PDS	Planetary Data System
PDS/CN	Planetary Data System / Central Node
PDS-D	Planetary Data System – Distribution System
PI	Principal Investigator
PIO	Public Information Office
PSG	Project Science Group
PSM	Project Science Manager
QA	Quality Assurance

RADIG	Radar Investigation Group
RDR	Reduced Data Record
SFOC	Space Flight Operations Center
SIS	Software Interface Specification
SPICE	Spacecraft ephemeris, Planet/satellite ephemeris, Instrument information, Camera orientation, Event information.
TBD	To be determined
TL	Team Leader
UDF	Universal Data Format
URL	Uniform Resource Locator
USGS	United States Geological Survey
XML	

Table A-1. Acronyms and abbreviations

APPENDIX B DEFINITION OF TERMS

The table below lists terms used in this document.

Archive	(1) One or more data sets stored for long-term preservation; (2) the logical structure of such an archive without regard to its physical location or the media on which it is stored.
Volume	One physical or logical unit of storage within an archive identified by a unique VOLUME_ID and VOLUME_NAME—for example, a single compact disk, magnetic tape. or disk partition. Depending on capacity of the volume relative to size of the data set(s), a single volume may hold several data sets or only part of one data set.
Volume Set	One or more volumes containing data with a common origin, time span, and/or theme and identified by a unique VOLUME_SET_ID and VOLUME_SET_NAME.
Volume Series	One or more volume sets containing data with a common origin, time span, and/or theme and identified by a unique VOLUME_SERIES_NAME.
Data Object	Data having a common origin or theme and organized into a recognized format. PDS recognizes about 30 objects including image, table, and text.
Label	A text file containing metadata (data about data) in an ODL keyword=value format which describes the structure and content of one or more data objects.
Data Product	A label and the one or more data objects that it describes.
Data Set	A grouping of data products having a common origin, time span, and/or theme and documentation sufficient to permit use without recourse to other data. A data set is identified by a unique DATA_SET_ID and DATA_SET_NAME.
Data Set Collection	A grouping of data sets which are related within a specific scientific area, practice, or objective (i.e., by observation type, discipline, target, or time) and identified by a unique DATA_SET_COLLECTION_ID and DATA_SET_COLLECTION_NAME.
Standard data product	A data product that has been defined during the proposal and selection process and that is contractually promised by the PI as part of the investigation. Standard data products are generated in a predefined way, using well-understood procedures, and processed in “pipeline” fashion.
Primary Data	The scientific measurements made by an instrument
Ancillary Data	Other data needed to understand or use the primary data. Examples include instrument settings, calibration data, files specifying observation geometry, processing software, and documents.

Table B-1. Definition of Terms

APPENDIX C ONLINE RESOURCES AND EXAMPLES

The following resources are available electronically either at the designated URL or through the *Archive Preparation Guide* web page:

http://pds.jpl.nasa.gov/documents/apg

C.1 PDS Web Pages

1. PDS Home Page (http://pds.jpl.nasa.gov/)

2. Web-based (quick-start introduction) to archiving data with PDS (http://pds.jpl.nasa.gov/qs/) (how is this different from the APG?)

3. *PDS Standards Reference* (JPL Document D-7669, Part 2) -Specific PDS data preparation standards for archive quality data sets. (http://pds.jpl.nasa.gov/stdref/)

4. *PDS Data Dictionary* - List of PDS-recognized keywords (and acceptable values where restrictions apply) (http://pds.jpl.nasa.gov/psdd.pdf)

C.2 PDS Nodes and Contact Information

PDS Personnel / Contact List – a table of PDS personnel and related contacts. These individuals are your most useful resource – contact them early and often.

http://pds.jpl.nasa.gov/data_services/contacts.html

C.3 Example Archiving Documents

1. Memorandum of Understanding (Not mentioned in APG. Isn't this really a mission-to-PDS issue? Do we want it here?)
   Mars Reconnaissance Orbiter: URL?

2. Data Management and Archiving Plan
   Messenger: URL?
   Deep Impact: URL?

   See also the separate Data Management Plan (URL?) and Archive Plan (URL?)

3. Interface Control Document/Software Interface Specification
   Huygens ICD: URL?
   Data Product SIS Template: URL? (is there an example document rather than template?)
   Archive Volume SIS Template: URL? (is there an example document rather than template?)

C.4 Example Catalog Files

1. CATINFO.TXT – CATINFO.TXT (representative) (URL?)
2. Mission Catalog File – MISSION.CAT (Voyager) (URL?)
3. Instrument Host Catalog File - INSTHOST.CAT (Voyager 1) (URL?)
4. Instrument Catalog File - INST.CAT (Voyager / RSS-VG1S) (URL?)
5. Data Set Catalog File - DATASET.CAT (Voyager 2 Triton Radio Occultation Reduced Data V1.0) (URL?)
6. Personnel Host Catalog File - PERSON.CAT (representative) (URL?)
7. References Catalog File - REF.CAT (representative) (URL?)
8. Target Catalog File - TARGET.CAT (Saturn) (URL?)

C.5 Example Product Files

1. IMAGE data object and label (Callisto)

   - 7000R.IMG (URL?)
   - 7000R.LBL (URL?)
   - VICAR2.TXT (URL?)
   - RLINEPRX.FMT (URL?)
   - RTLMTAB.FMT (URL?)

2. TABLE object and label

   - ??.TAB (URL?)
   - ??.LBL (URL?)

3. INDEX_TABLE and label (index and cumulative index files)

- IMGINDEX.LBL (URL?)
- IMGINDEX.TAB (URL?)
- CUMINDEX.LBL (URL?)
- CUMINDEX.TAB (URL?)

C.6 Example Calibration Files / Algorithms

TBD (URL?)

C.7 Example Data Set Citation descriptions

1. Examples from Mars Exploration Rover (MER) (URL?)

```
DATA_SET_ID      = "MER2-M-MTES-3-RDR-V1.0"
DATA_SET_NAME    = "MER 2 MARS MINI-TES RDR V1.0"
CITATION_DESC    = "Christensen, Phil, 'MER 2 Mars
                    Mini-TES RDR V1.0', NASA Planetary
                    Data System,
                    MER2-M-MTES-3-RDR-V1.0, 2004."


DATA_SET_ID      = "MER1-M-MTES-2-EDR-V1.0"
DATA_SET_NAME    = "MER 1 MARS MINI-TES EDR V1.0"
CITATION_DESC    = "Christensen, Phil, 'MER 1 Mars
                    Mini-Thermal Emission Spectrometer
                    EDR V1.0', NASA Planetary Data System,
                    MER1-M-MTES-2-EDR-V1.0, 2004."
```

C.8 PDS Tool Suite

The latest set of released PDS Tools may be downloaded from the PDS Software Library at: http://pds.jpl.nasa.gov/software.html. The tools may also available from the following ftp site: ftp://pds.jpl.nasa.gov/tools/PDS-latest.

APPENDIX D INTERFACE CONTROL DOCUMENT (ICD) OUTLINE

The overall structure and content of the ICD/SIS should reflect the following:

1	INTRODUCTION
1.1	PURPOSE AND SCOPE
1.2	CONTENTS
1.3	INTENDED READERSHIP
1.4	APPLICABLE DOCUMENTS
1.5	RELATIONSHIPS TO OTHER INTERFACES
1.6	ACRONYMS AND ABBREVIATIONS
1.7	CONTACT NAMES AND ADDRESSES
2	OVERVIEW OF PROCESS AND PRODUCT GENERATION
3	ARCHIVE FORMAT AND CONTENT
3.1	FORMAT
3.1.1	Volume Format
3.1.2	Data Set Format
3.1.3	File Formats
3.2	STANDARDS USED IN DATA PRODUCT GENERATION
3.2.1	PDS Standards
3.2.2	Time Standards
3.3.3	Coordinate Systems
3.3	DATA VALIDATION
3.4	CONTENT
3.4.1	Volume Set
3.4.2	Data Set
3.4.3	Directories
3.4.3.1	Root Directory
3.4.3.2	Calibration Directory
3.4.3.3	Catalog Directory
3.4.3.3.1	Mission Catalog File to be copied from ESA]
3.4.3.3.2	Mission Host Catalog File to be copied from ESA]
3.4.3.3.3	Host Instrument Catalog File
3.4.3.3.4	Instrument Catalog File
3.4.3.3.5	Personnel Catalog File
3.4.3.3.6	Reference Catalog File
3.4.3.3.7	Dataset Catalog File
3.4.3.3.8	Map Projections Catalog File
3.4.3.3.9	Targets Catalog File
3.4.3.4	Index Directory
3.4.3.4.1	Dataset Index File, index.lbl and index.tab
3.4.3.4.2	Geometric Index File, geoindex.lbl and geoindex.tab
3.4.3.4.3	other Index Files
3.4.3.5	Browse Directory and Browse Files
3.4.3.5	Geometry Directory
3.4.3.6	Software Directory
3.4.3.7	Gazetter Directory
3.4.3.8	Label Directory
3.4.3.9	Document Directory
3.4.3.10	Extras Directory
3.4.3.11	Data Directory
3.4.4	Data and Label Files

APPENDIX E CODMAC LEVELS

This section needs work. NASA and CODMAC levels do not match perfectly. Is this the CODMAC list with NASA shoe-horned in, vice versa, or something else? What are we trying to say here; what is needed to support the APG text? Having the same numbered levels (e.g., Level 3) in both columns will be confusing. What is CODMAC?

NASA	CODMAC	Description
Packet data	Raw - Level 1	Telemetry data stream as received at the ground station, with science and engineering data embedded.
Level-0	Edited - Level 2	Instrument science data (e.g., raw voltages, counts) at full resolution, time ordered, with duplicates and transmission errors removed.
Level 1-A	Calibrated - Level 3	Level 0 data that have been located in space and may have been transformed (e.g., calibrated, rearranged) in a reversible manner and packaged with needed ancillary ~~and auxiliary~~ data (e.g., radiances with the calibration equations applied).
Level 1-B	Resampled - Level 4	Irreversibly transformed (e.g., resampled, remapped, calibrated) values of the instrument measurements (e.g., radiances, magnetic field strength).
Level 1-C	Derived - Level 5	Level 1A or 1B data that have been resampled and mapped onto uniform space-time grids. The data are calibrated (i.e., radiometrically corrected) and may have additional corrections applied (e.g., terrain correction).
Level 2	Derived - Level 5	Geophysical parameters, generally derived from Level 1 data, and located in space and time commensurate with instrument location, pointing, and sampling.
Level 3	Derived - Level 5	Geophysical parameters mapped onto uniform space-time grids.

Table E-1. NASA and CODMAC Definitions of Processing Levels for Science Data Sets

APPENDIX F DN'S, HOST INSTITUTIONS, AREAS OF INTEREST

TBW